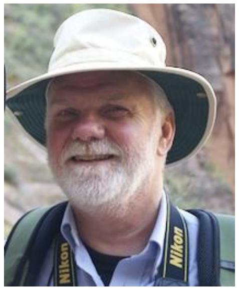
Distinguished R.L. Anderson Speakers

Dr. Simon Sheather
Dean of the Gatton College of Business and Economics.
University of Kentucky

April 19, 2019

Data Scientists and Statisticians: Competitors or Collaborators?
In this talk we examine the fields of data science and statistics highlighting their commonalities and their differences. In particular, we describe areas where statistical methods are of importance in data science as well as the reverse. We also discuss areas which can be best described as “Much ado about the wrong thing”. Both personal views and those from the literature will be provided.



Dr. Douglas Bates
Emeritus Professor,
Department of Statistics,
University of Wisconsin-Madison

April 20, 2018

The Evolution of Languages for Data Analysis
I recently realized that it has been 50 years since I had my first (and only) course in computer science and my first (and only) undergraduate course in statistics. Well, there have been a few changes during those 50 years. I have been fortunate to participate in some of the development of the S language and later the development of R. I was also an early user of Python. For the last 5 years I have been developing Julia packages. In this talk I will look back on the evolution of languages for data analysis and offer some suggestions on where we might expect the field to go.

Dr. Michael Kutner
Professor, Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

April 14, 2017

Statistical Tests for Interactions and Main Effects in Two-Factor Unbalanced Cross-Classified Fixed Effects Experiments with No Data in Some Cells
First, I will give some historical remarks regarding the career of R. L. (Dick) Anderson. Then I will discuss the statistical tests of both the interaction effects and main effects in unbalanced two-factor cross-classified fixed effects experiments with no data in some cells. During the latter part of Dick Anderson’s highly regarded career the testing was problematic when using standard statistical software packages default options such as SAS, SPSS and BMDP. This lecture will lay out the fundamental reasons why these tests are

			not routine and will provide you with how one can generate tests for interactions and main effects that un-fortunately are dependent upon which cells have no data. Even today, very few textbooks and/or classes in analysis of variance handle this topic particularly well. Furthermore, the default options provided by SAS, SPSS and BMDP are still problematic.
	Dr. Nan Laird Harvey V. Fineberg Research Professor of Biostatistics at the Harvard T.H. Chen School of Public Health	April 8, 2016	**Multivariate Problems in the Genetic Analysis of Complex Disease** Complex diseases have multiple underlying contributing factors, both genetic and environmental. In addition, the disease syndrome is often characterized by measured clinical traits that may be analyzed for association with genes along with the disease status. Genome Wide Association Analysis (GWAS) has been highly successful in identifying some genetic loci associated with many disease syndromes and/or selected traits. The purpose of the analysis of multiple traits may be to show consistency and thereby strengthen the evidence, or to identify different loci for different traits, or to gain additional power for new loci. In this talk we describe an approach to integrating multiple phenotypes based on the concepts of heritability and co-heritability. Our approach is designed for GWAS and uses the genetic data both for the estimation of heritability and using samples of cases and controls and for testing association.
	Dr. Dennis Cook Professor, School of Statistics, University of Minnesota	Feb. 26, 2015	**Envelopes: A Novel Class of Methods for Multivariate Statistics** An envelope is a nascent construct for increasing efficiency in multivariate statistics without altering the traditional goals. Envelope estimators have the potential to be substantially less variable than standard estimators, sometimes equivalent to taking thousands of additional observations. Improvements in efficiency are made possible by recognizing that the data may contain variation that is effectively immaterial to estimation. This informal notion leads to a general construct – an envelope – for enveloping the material information and thereby reducing estimative variation and improving inference. Envelopes also link with some standard multivariate methodology. For instance, it was recently discovered that partial least squares regression

depends fundamentally on an envelope and this envelope can be used as a well-defined parameter that characterizes partial least squares. The establishment of an envelope as the nucleus of partial least squares then opens the door to pursuing the same goals but using envelope estimators that can significantly improve upon partial least squares predictions.

We will begin with an intuitive introduction to response envelopes in the context of multivariate linear regression and then briefly describe some of their inner workings. This will be followed by a discussion of predictor envelopes and their connection to partial least squares. We will also describe briefly how to extend the scope of envelope methods well beyond linear models. The discussion will include several examples for illustration. Emphasis will be placed on concept and their potential impact on data analysis.

Dr. Stephen Fienberg
Maurice Falk University, Professor of Statistics and Social Science, Carnegie Mellon University

April 18, 2014

Multivariate Problems in the Genetic Analysis of Complex Disease
What makes "big data" big? The number of people under observation? The number of data elements per person? The complexities of dependencies among the units and data elements? However we choose to describe big data, most people agree that it often arises from the extensive digital traces we produce from virtually all facets of our lives. The Living Analytics paradigm involves a cycle of learning through experimentation in networked environments which poses new challenges for multivariate statistical research. I will describe two of these in the context of the projects of the joint Carnegie Mellon-Singapore Management University Living Analytics Research Centre: (1) the need for models describing individual trajectories in time and space, and (2) the role of experiments in networked environments. A third challenge is that posed by requirements to protect the privacy of individuals whose data are used in the context of living analytics research. I will focus on aspects of extending the now standard statistical and cryptographic approaches to the privacy to networked data and I will review some progress on the topic to date.

Dr. Gary Koch
Department of Biostatistics,
University of North Carolina,
Chapel Hill

April 24,
2013

Analysis of Covariance: Model-based and Nonparametric

For randomized clinical trials with at least moderate sample size, adjustment of comparisons between treatments for baseline covariables can be helpful for two reasons. One is enhancement of power, and the other is the removal of the influence of baseline imbalances for the covariables. Adjustment for baseline covariables can either be through generalized linear (or semi-parametric) models or through a nonparametric extension of Mantel-Haenszel methods. The former has the limitation of assumptions that may be debatable or unrealistic, although it can have the advantage of fully describing the relationship of an endpoint to both treatments and covariables in a general population. The latter has the advantage of no external assumptions (beyond its intrinsic assumptions of valid randomization and valid data), although it only enables inference for the comparison between treatments for the randomized population. The nonparametric method has invocation by constraining differences between treatments for means of covariables to 0 in a multivariate vector that additionally includes the unadjusted treatment effect sizes for the endpoints under assessment. Such nonparametric randomization based analysis of covariance (RBANCOVA) is applicable to differences between means for continuous measurements (or their ranks), differences between proportions, log hazard ratios for time to event data, log incidence density ratios for counted event data, and rank measures of association for ordinal data. Also, extensions to account for stratification factors in the randomization are available as well. Several examples which illustrate RBANCOVA and model based counterparts have discussion.

	Dr. Frank Harrell, Jr. Professor, Department of Biostatistics, Vanderbilt University School of Medicine	April 23, 2012	**Statisticians, persons interested in personalized medicine, biomarkers, reproducible research, or clinical epidemiology** There are many ways to personalize the diagnosis and treatment of diseases, pharmacogenomics being one of them. Personalization can be based on routinely collected information, molecular signatures, or on repeated trials on the patient whose treatment plan is being devised. However, current emphases in personalized medicine research often ignore characteristics known to impact treatment benefit, in favor of tests that either generate more revenue or are developed with research that is perhaps easier to fund than "low-tech" research. Failure of the research community to fully utilize rich datasets generated by randomized clinical trials only hightens this concern. Research supporting personalized medicine can be made more rigorous and relevant. For example in acute diseases, multi-period crossover studies can be used to measure individual response to therapy, and these studies can provide an upper bound on the genome by treatment interaction. When patient by treatment interaction is demonstrated, crossover studies can form an ideal basis for pharmacogenomics. However, even with the best within-patient data, group average treatment effects need to be incorporated in order for predictions for individual patients to have high precision. There are a few ways to do personalized medicine well but a multitude of ways to do it poorly. Biomarker research in particular has not fulfilled its early promises, a major reason being flawed methodology. The flaws include faulty experimental design, bias, overfitting, weak validation, irreproducible research, data processing and analysis practices, and failure to rigorously show that the new markers add information to readily available clinical data. This will be discussed in terms of Platt's concept of "strong inference", seeking alternative explanations of findings, and sensitivity analysis. This talk is also a call for the biostatistics and clinical epidemiology communities to be more integrally involved in research related to personalized medicine

	Dr. Michael I. Jordan Distinguished Professor of Statistics & Computer Science, University of California, Berkeley	April 7, 2011	**Statistical Inference of Protein Structure and Function** The study of the structure and function of proteins raises many problems that offer challenges and opportunities for statistical research. I will overview my experiences in several such problem domains, ranging from domains where off-the-shelf ideas can be fruitfully applied to domains that require new thinking. These are: (1) the identification of active sites in enzymes; (2) the modeling of protein backbone configurations; (3) the prediction of molecular function based on phylogeny; (4) joint inference of alignment and phylogeny.